KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

The Gaussian rank correlation estimator: robustness properties

Kris Boudt, Jonathan Cornelissen and Christophe Croux

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 1023

Noname manuscript No.
(will be inserted by the editor)

The Gaussian rank correlation estimator: robustness properties

Kris Boudt · Jonathan Cornelissen · Christophe Croux

Received: date / Accepted: date

Abstract The Gaussian rank correlation equals the usual correlation coefficient computed from the normal scores of the data. Although its influence function is unbounded, it still has attractive robustness properties. In particular, its breakdown point is above 12%. Moreover, the estimator is consistent and asymptotically efficient at the normal distribution. The correlation matrix based on the Gaussian rank correlation is always positive semidefinite, and very easy to compute, also in high dimensions. A simulation study confirms the good efficiency and robustness properties of the proposed estimator with respect to the popular Kendall and Spearman correlation measures. In the empirical application, we show how it can be used for multivariate outlier detection based on robust principal component analysis.

Keywords Breakdown · Correlation · Efficiency · Robustness · Van der Waerden

Financial support from the Flemish IWT (Institute for Science and Innovation) is gratefully acknowledged.

Kris Boudt
Lessius University College, Antwerp, Belgium,
and K.U.Leuven, Faculty of Business and Economics, Belgium.
Tel.: +32-03-2011810
E-mail: kris.boudt@econ.kuleuven.be

Jonathan Cornelissen
K.U.Leuven, Faculty of Business and Economics, Belgium.
Tel.: +32-16-326728
E-mail: jonathan.cornelissen@econ.kuleuven.be

Christophe Croux
K.U.Leuven, Faculty of Business and Economics, Belgium.
Tel.: +32-16-326958
E-mail: christophe.croux@econ.kuleuven.be

1 Introduction

It is well known that the value of the classical Pearson correlation estimator can be highly affected by the presence of only a small amount of outliers. To overcome this sensitivity to outliers, estimators based on ranks, such as the Spearman and Kendall correlation, can be used (see e.g. Croux and Dehon (2010) for a recent review). Correlation matrices constructed from the Spearman or Kendall correlation are robust and very fast to compute, also in high dimensions. They have been applied in robust principal component analysis (Van Aelst et al., 2010), robust data mining (Alqallaf et al., 2002), and robust least angle regression (Khan et al., 2007), among others. However, to obtain consistency at the normal distribution, one needs to apply a transformation, and the resulting correlation matrix is no longer guaranteed to be positive semidefinite.

We propose a positive semidefinite correlation estimator based on ranks that is consistent and asymptotically efficient at the normal distribution. Moreover, it has a quite high robustness to outliers at finite samples. We call this estimator the Gaussian rank correlation. For a bivariate sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, it is constructed by first computing the ranks of each observation. Denote by $R(x_i)$ and $R(y_i)$ the rank of x_i and y_i respectively, for $1 \leq i \leq n$. Next, the corresponding Gaussian (or normal) scores are obtained by plugging these ranks in the quantile function Φ^{-1} of the standard normal distribution. The Gaussian rank correlation (GRCor) is then the conventional correlation computed from these scores:

$$\text{GRCor} = \frac{\sum_{i=1}^{n} \Phi^{-1}(\frac{R(x_i)}{n+1})\Phi^{-1}(\frac{R(y_i)}{n+1})}{\sqrt{\sum_{i=1}^{n} \Phi^{-1}(\frac{R(x_i)}{n+1})^2 \sum_{i=1}^{n} \Phi^{-1}(\frac{R(y_i)}{n+1})^2}}. \tag{1.1}$$

Note that the value of the denominator only depends on the sample size n and not on the data. The transformation of data to the Gaussian scores, also called the Van Der Waerden scores or the normal scores, to obtain correlation estimators is not new in the statistical literature (see e.g. Hájek and Sidak (1967)). However, a study of the properties of the correlation estimator (1.1) from a robustness perspective has been lacking in the literature, up to the knowledge of the authors.

The robustness of the GRCor follows from the use of ranks. Although the estimator has an unbounded influence function, its breakdown point is above 12.4 %. An important advantage of the proposed estimator is that at the normal distribution, it is asymptotically as efficient as the sample correlation coefficient. Moreover, no transformation is needed to obtain consistency for the correlation coefficient of a bivariate normal distribution. By consequence, we can construct an estimate for the correlation matrix of a multivariate normal distribution by estimating each element of this matrix by the GRCor. The resulting matrix estimate is ensured to be positive semidefinite, since the GRCor matrix is the standard correlation matrix computed on the data transformed into the Van Der Waerden scores.

In Sections 2 and 3 we investigate the robustness properties of the GRCor, by computing its breakdown point, sensitivity curve and influence function. Subsequently, the performance of the Gaussian rank estimator is compared with the Kendall and Spearman correlation estimators in a simulation study. The usefulness of the GRCor on real data is illustrated in Section 5, where outliers are detected using principal component analysis of the GRCor correlation matrix. The last section summarizes our main findings.

2 Breakdown point

The breakdown point of an estimator is the smallest fraction of data contamination that can make the estimator uninformative. For correlation estimates, it is especially interesting to study the contamination needed to invert the sign of the correlation estimate. More formally, for a sample $Z_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ and a correlation estimator $\hat{\rho}$, we define the finite sample breakdown point towards zero as

$$\varepsilon_n(\hat{\rho}; Z_n) = \min_k \{\frac{k}{n} : \inf_{Z_n^k} \hat{\rho}(Z_n^k)\, \hat{\rho}(Z_n) \leq 0\},$$

where Z_n^k is obtained by replacing any k observations of Z_n by arbitrary values. This finite sample breakpoint strongly depends on the sample configuration. We focus on a sample S_n with identical component observations, namely $S_n = \{(x_1, x_1), \ldots, (x_n, x_n)\}$. We assume, and this holds for all estimators we consider, that $\hat{\rho}(S_n) = 1$. The finite sample breakdown point of a correlation estimator $\hat{\rho}$ is then defined as

$$\varepsilon_n(\hat{\rho}) := \varepsilon_n(\hat{\rho}; S_n) = \min_k \{\frac{k}{n} : \inf_{S_n^k} \hat{\rho}(S_n^k) \leq 0\}. \tag{2.1}$$

It equals the smallest number of observations one needs to replace from a perfectly correlated bivariate sample, to make the correlation coefficient become negative. A similar definition of breakdown point was used by Capéraà and Guillem (1997) in the context of testing for no correlation, and by Grize (1978). Note that there is no canonical definition of breakdown point for correlation estimators (see the discussion and rejoinder in Davies and Gather (2005)).

Proposition 1 describes the type of outliers that induces the largest downward bias in the GRCor estimate at the sample S_n. It is a reformulation of Proposition 2.3 of Capéraà and Guillem (1997), where the proof can be found.

Proposition 1 *Consider the sample S_n where the component observations are identical. Assume without loss of generality that $x_1 < \ldots < x_n$. For every $1 \leq k \leq n$, and for every $0 \leq r \leq k$, denote by $S_n^{k,r}$ the sample where k elements of S_n are replaced as follows:*

$$\begin{cases} (y_1, \ldots, y_{k-r}) \text{ by } (y'_1, \ldots, y'_{k-r}) & \text{with } y'_1 > \ldots > y'_{k-r} > \max_i y_i \text{ (if } k > r) \\ (y_{n-r+1}, \ldots, y_n) \text{ by } (y'_{n-r+1}, \ldots, y'_n) & \text{with } y'_n < \ldots < y'_{n-r+1} < \min_i y_i \text{ (if } r > 0). \end{cases}$$

Let S_n^k be any sample where k elements of S_n are replaced. Then it holds that

$$\inf_{0 \leq r \leq k} \hat{\rho}(S_n^{k,r}) \leq \inf_{S_n^k} \hat{\rho}(S_n^k).$$

For computing the breakdown point (2.1) of the GRCor estimator, we proceed as follows. For a given sample size n, and for $0 < k < n$, we compute the value r^* yielding the minimal $\hat{\rho}(S_n^{k,r})$, with $0 \leq r \leq k$. Denote $S_n^{k^*} = S_n^{k,r^*}$. It turns out, as we verified numerically for sample sizes up to $n = 10000$, that $r^* = \lfloor k/2 \rfloor$, yielding the expression

$$\hat{\rho}(S_n^{k*}) = c_n \begin{cases} \sum_{i=1+k/2}^{n-k/2} z_i^2 - 2\sum_{i=1}^{k/2} z_i^2 & \text{if } k \text{ is even} \\ \sum_{i=2+\lfloor k/2 \rfloor}^{n-\lfloor k/2 \rfloor} z_i z_{i-1} - 2\sum_{i=1}^{\lfloor k/2 \rfloor} z_i^2 - z_{\lfloor k/2 \rfloor + 1}^2 & \text{if } k \text{ is odd,} \end{cases} \tag{2.2}$$

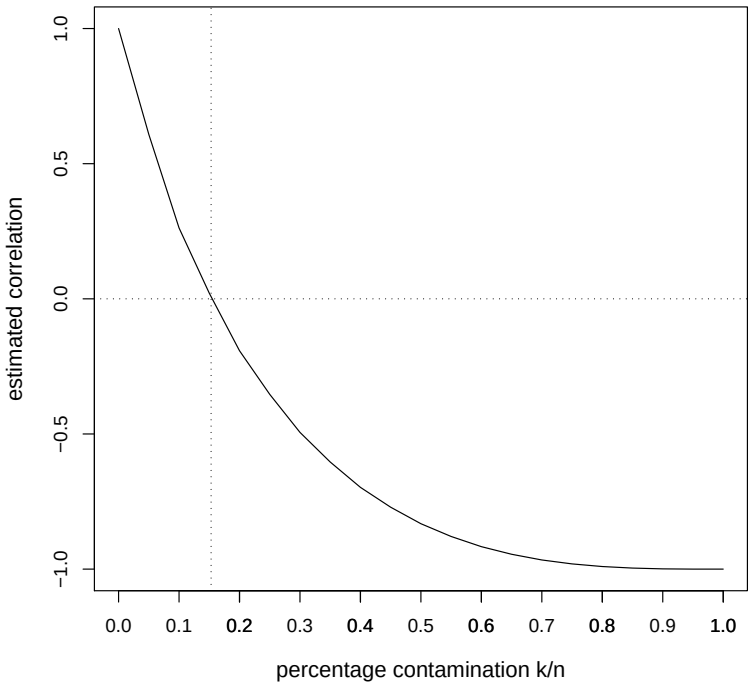

Fig. 1: Value of $\hat{\rho}(S_n^{k^*})$ versus the percentage of contamination k/n, with $n = 20$.

with $z_i = \Phi^{-1}(i/(n+1))$ and $c_n = 1/\sum_{i=1}^{n} z_i^2$. Figure 1 plots $\hat{\rho}(S_n^{k^*})$ versus the percentage of outliers k/n for $n = 20$. We observe that the GRCor decreases as the percentage of outliers increases, as expected. The finite sample breakdown point for this value of n is the smallest value of k/n yielding a non-positive correlation. As can be seen from Figure 1, this occurs if k/n is close to 15%.

In Figure 2, we present the finite sample breakdown point $\varepsilon_n(\hat{\rho})$ of the GRCov as a function of the sample size. We observe that the breakdown point for finite samples remains above 12%. For very small sample sizes, below $n = 20$, the finite sample breakdown point is even considerably higher. In the next proposition we give the asymptotic value of $\varepsilon_n(\hat{\rho})$ (the derivation is in Appendix).

Proposition 2 *Let $\varepsilon^* = \lim_{n\to\infty} \varepsilon_n(\hat{\rho})$. For the GRCor estimator it holds that ε^* is the unique solution of the equation*

$$\varepsilon/2 - \Phi^{-1}(\varepsilon/2)\phi(\Phi^{-1}(\varepsilon/2)) = \frac{1}{4}, \tag{2.3}$$

with $0 < \varepsilon < 1$.

Solving the equation (2.3) numerically yields an asymptotic breakdown point of $\varepsilon^* = 0.124$ for the Gaussian rank correlation. The finite sample breakdown

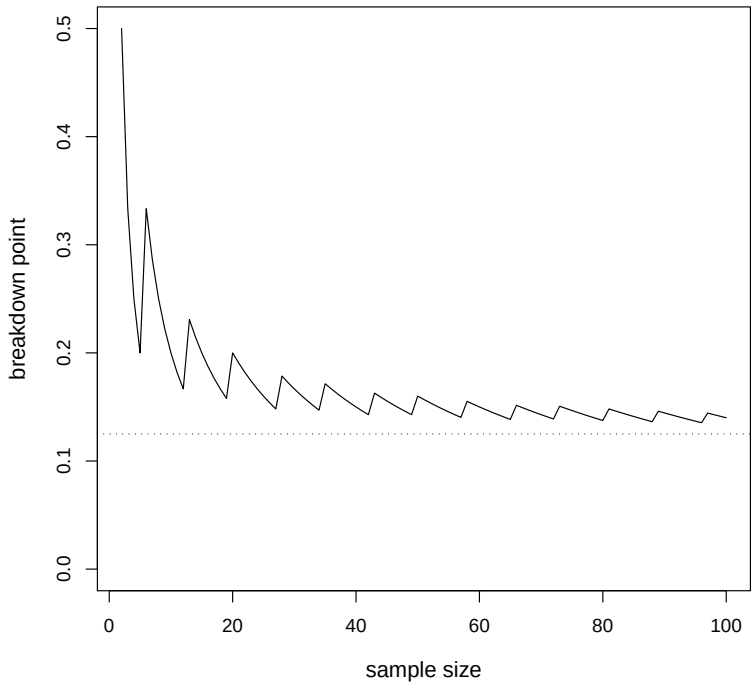

Fig. 2: Finite sample breakdown points of the GRCor as a function of the sample size. The dashed line corresponds to the asymptotic breakdown point of 12.4%

point of the standard Pearson correlation is only $1/n$, resulting in $\varepsilon^* = 0$. For the Spearman correlation, one has $\varepsilon^* = 0.206$, and for the Kendall correlation $\varepsilon^* = 0.293$ (e.g. Grize (1978), and Davies and Gather (2005)), showing that they are more robust to large amounts of outliers than the Gaussian rank correlation.

3 Sensitivity curve

The sensitivity curve measures the robustness of an estimator to small amounts of contamination. For a bivariate sample $Z_{n-1} = \{(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\}$ it is defined by

$$SC_n(x, y) = n[\hat{\rho}(Z_{n-1} \cup \{(x, y)\}) - \hat{\rho}(Z_{n-1})]. \tag{3.1}$$

It measures the change in the estimator caused by adding (x, y) to the clean sample, standardized by the amount of contamination. Figure 3 pictures the sensitivity curve of the GRCor estimator for $n = 200$, averaged over 100 random samples of 199 observations from a bivariate normal distribution with correlation $\rho = 0.8$. We see that the bias induced by adding one outlying couple (x, y) is the largest when

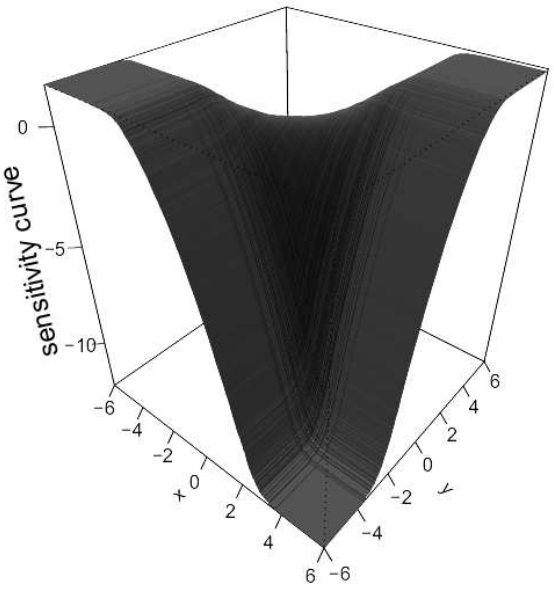

Fig. 3: Sensitivity curve averaged over 100 samples of 199 observations from a standard bivariate normal distribution with correlation 0.8.

that observation is outlying both with respect to the correlation structure as well as to the marginal distributions, i.e. for a large positive x and a large negative y, or for a large negative x and a large positive y.

Figure 4 shows the sensitivity curve $SC_n(x, -x)$ as a function of x, for the GRCor and classical Pearson correlation. Both are decreasing functions of $|x|$, but the lower bound for the GRCor is much smaller than for the Pearson correlation, due to the use of ranks. Although these sensitivity curves are bounded for finite n, this no longer holds when n tends to infinity.

To focus ideas, consider $Z_{n-1} = S_{n-1}$ and denote

$$\gamma_n^* = \sup_{x,y} |SC_n(x, y)|,$$

where the sensitivity curve SC_n is computed at S_{n-1}. The limit quantity $\gamma^* = \lim_{n\to\infty} \gamma_n^*$ measures the gross-error sensitivity of the estimator. For the Pearson correlation estimator, one has that $\gamma_n^* = 2n$, and the gross-error sensitivity is infinite. The next proposition (proof in Appendix) shows that the Gaussian rank correlation is much more robust with respect to single outliers.

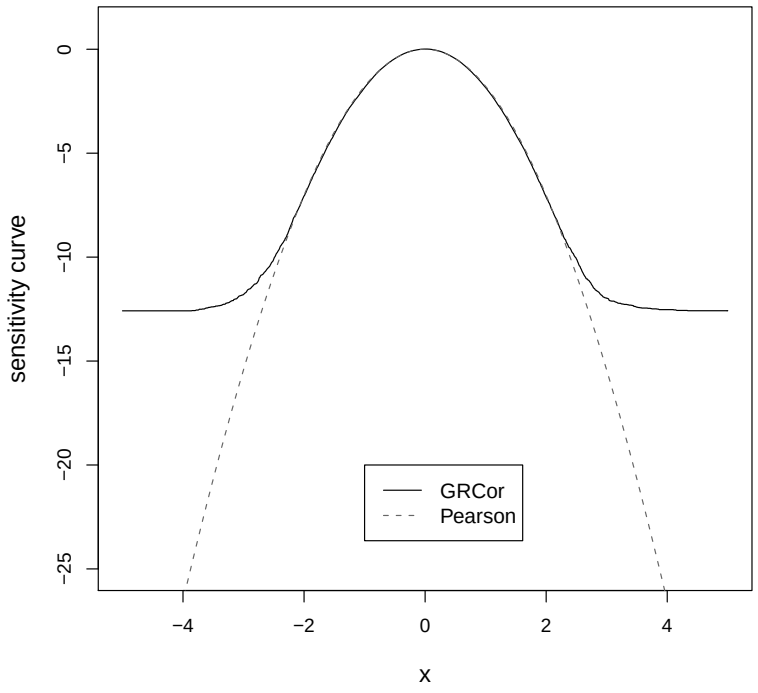

Fig. 4: Sensitivity curve $SC_n(x, -x)$ as a function of x, averaged over 100 samples of 199 observations observations from a standard bivariate normal distribution with correlation 0.8.

Proposition 3 *The gross error sensitivity of the GRCor is infinite but the divergence to infinity is only at a logarithmic rate:*

$$\gamma_n^* \sim 2\log(n).$$

An alternative way to quantify the effect of a small amount of contamination on the GRCor, is the influence function (e.g. Maronna et al. (2006)). To compute it, we need a definition of the GRCor at the population level. Assume that the bivariate random variable (X, Y) follows a distribution H. The influence function (IF) of a statistical functional T at a distribution H is defined as

$$\mathrm{IF}((x_0, y_0), T, H) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)H + \varepsilon \Delta_{(x_0,y_0)}) - T(H)}{\varepsilon} \tag{3.2}$$

where $\Delta_{(x_0,y_0)}$ is a Dirac measure putting all its mass at (x_0, y_0). It can be interpreted as the effect that an infinitesimal amount of contamination placed at (x_0, y_0) has on T, for data coming from the distribution H. Note that the influence function is defined at the population level, and that the IF of an estimator

refers to the IF of the associated functional representation of the estimator. For the sample (X, Y) coming from the arbitrary distribution H, with marginal distribution F and G, the statistical functional associated with the GRCor is given by

$$\text{GRCor}(H) = \int \Phi^{-1}(F(x))\ \Phi^{-1}(G(y))\ dH(x, y). \tag{3.3}$$

Given this functional representation of the GRCor, we now present the influence function at the bivariate normal distribution. A proof is provided in the Appendix.

Proposition 4 *The influence function of the GRCor at the bivariate normal distribution Φ_ρ, having correlation coefficient ρ, is given by*

$$IF((x_0, y_0), GRCor, \Phi_\rho) = x_0 y_0 - \rho(\frac{x_0^2 + y_0^2}{2}). \tag{3.4}$$

The IF is thus unbounded, and is exactly the same as the IF of the Pearson correlation (Devlin et al., 1975). However, this result is somehow misleading, since it is based on the asymptotic representation of the estimator. At finite samples, the GRCor is much less sensitivity to outliers than the Pearson correlation, as we showed in Proposition 3. This finding will be confirmed in the simulation study presented in the next section.

4 Simulation study

By means of a simulation study, we compare the finite sample performance of several correlation estimators, both in the absence and presence of outliers. Apart from the GRCor, we consider the Pearson, Spearman, Kendall correlation estimators and the correlation estimator associated with the Minimum Covariance Determinant (MCD). First, we focus on the bivariate case, where we assess the performance of the different estimators in terms of mean squared error. Second, we study the multivariate case, where we also investigate the positive semidefiniteness of the correlation matrix estimates.

Competing correlation estimators: For a bivariate sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ the classical Pearson's estimator of correlation is given by

$$R_P = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \sum_{i=1}^n (y_i - \overline{y})}},$$

where $\overline{x}$ and $\overline{y}$ are the sample means. The most popular alternative for the Pearson correlation estimator is the Spearman's rank correlation r_S (Spearman, 1904), defined as the Pearson correlation between the ranks of the observations. A consistent version of this estimator for the population correlation ρ of a bivariate normal distribution is obtained by the transformation

$$R_S = 2\sin(\frac{1}{6}\pi\, r_S).$$

Another nonparametric correlation measure is Kendall's correlation (Kendall, 1938), defined as

$$r_K = \frac{2}{n(n-1)} \sum_{i<j} \text{sign}((x_i - x_j)(y_i - y_j)).$$

To obtain a consistent version at the normal distribution, we apply the transformation

$$R_K = \sin(\frac{1}{2}\pi\, r_K).$$

Recall that no transformation is needed for the Pearson and GRCor correlation estimators. Finally, we compare with the Minimum Covariance Determinant estimator (MCD, Rousseeuw and Van Driessen (1999)) with 50% breakdown point and additional reweighting step, and compute the associated correlation matrix. In the bivariate case we get

$$R_{MCD} = \frac{\hat{C}_{12}}{\sqrt{\hat{C}_{11}\hat{C}_{22}}},$$

with $\hat{C}$ the MCD covariance matrix estimator. We use the R-command covMcd from the robustbase package, with default options, for computing the MCD (Rousseeuw et al., 2009). The computation of the MCD is much more time consuming than for the GRCor correlation.

Bivariate simulation design: We generate $m = 1000$ samples of size $n = 50, 100, 200$ from a bivariate normal with correlation coefficient $\rho = 0.2$ or $\rho = 0.8$ (simulations for other values of ρ result in similar conclusions). We introduce outliers in the data by replacing a percentage ε of the observations by the the point $(5, -5)$, where the sensitivity curve of the GRCor is close to its most extreme value, see Figure 3. For each sample j, the correlation coefficient is estimated by $\hat{\rho}_j$, and the Mean Squared Error (MSE) is computed as

$$\text{MSE} = \frac{1}{m}\sum_{j=1}^{m}(\hat{\rho}_j - \rho)^2. \tag{4.1}$$

Table 1 reports the MSE for the different estimators considered. Standard errors around the reported results are about 2% of the respective results.

In the absence of outliers, we find that the Pearson correlation has the highest precision, as expected. The GRCor comes out second, followed by Kendall and Spearman correlation. The correlation estimates associated with the MCD covariance estimator are clearly less precise regardless of the sample size or the correlation coefficient. For all estimators, we find that the precision is higher in case the true correlation coefficient ρ is closer to one. At the bottom of Table 1 we report the asymptotic variance of the respective estimators, as documented in Croux and Dehon (2010). The asymptotic variance of the Gaussian rank correlation is equal to that of the Pearson correlation.

In the presence of outliers, we notice that even in case of a small percentage of outliers the Pearson correlation becomes very unprecise. Focussing on the robust correlation estimators, we observe a different behavior for mild contamination, i.e. $\varepsilon = 1\%$, and more pronounced contamination, i.e. $\varepsilon = 10\%$. In case of mild contamination, the most precise estimator is most often the Kendall correlation. For $\varepsilon = 10\%$, the MCD estimator becomes by a wide margin the estimator with the smallest MSE. The GRCor is the least precise of all considered robust estimators in case of outliers, but still far more precise than the Pearson correlation. Moreover, in higher dimensions the GRCor has the advantage of being positive definite, in contrast to estimates based on R_S and R_K, while keeping its robustness.

Table 1: Simulated MSE (multiplied by the sample size) of several correlation estimators at a bivariate standard normal distribution with $\rho = 0.2$ and $\rho = 0.8$, for sample sizes $n = 50, 100, 200$ and a fraction ε of outliers at position (5,-5).

	$\rho = 0.2$				$\rho = 0.8$			
n*MSE	$\varepsilon = 0\%$	$\varepsilon = 1\%$	$\varepsilon = 5\%$	$\varepsilon = 10\%$	$\varepsilon = 0\%$	$\varepsilon = 1\%$	$\varepsilon = 5\%$	$\varepsilon = 10\%$
$n = 50$								
Pearson	0.95	9.07	26.89	37.56	0.14	19.81	60.44	84.46
Spearman	1.04	1.24	3.04	6.40	0.20	0.83	4.92	12.28
Kendall	1.06	1.28	2.96	6.12	0.18	0.44	2.37	6.49
GRCor	0.95	1.59	4.57	8.30	0.18	2.08	9.50	18.32
MCD	2.92	2.78	2.57	2.31	0.54	0.49	0.45	0.40
$n = 100$								
Pearson	0.93	6.68	45.54	74.22	0.14	13.93	102.30	167.14
Spearman	1.03	1.13	4.05	12.00	0.18	0.50	6.72	23.99
Kendall	1.04	1.15	3.83	11.16	0.17	0.29	3.03	12.41
GRCor	0.93	1.39	6.98	15.96	0.16	1.49	14.92	36.32
MCD	2.70	2.62	2.33	2.03	0.44	0.42	0.35	0.32
$n = 200$								
Pearson	0.92	12.47	90.00	147.76	0.13	26.81	202.07	332.83
Spearman	1.01	1.27	7.15	23.38	0.17	0.75	13.01	47.48
Kendall	1.02	1.26	6.58	21.49	0.15	0.37	5.71	24.36
GRCor	0.92	2.03	13.37	31.35	0.14	2.94	29.91	72.31
MCD	2.50	2.49	2.20	1.95	0.35	0.36	0.33	0.28

For $\rho = 0.2$ (0.8), the asymptotic variance of Pearson, Spearman, Kendall, GRCor and MCD is 0.92 (0.13), 1.02 (0.16), 1.01 (0.15), 0.92 (0.13) and 2.26 (0.32) respectively.

Multivariate simulation design: We now generate 1000 samples from a 10 dimensional multivariate normal distribution with mean zero. Denote by Σ the corresponding covariance matrix, for which all diagonal elements are equal to 1 and all off-diagonal elements equal to ρ. As in Branco et al. (2005), we consider both symmetric and asymmetric contamination. In case of symmetric contamination, a fraction ε of the sample follows a multivariate normal distribution with covariance matrix $5 * \Sigma$. In case of asymmetric contamination, a fraction ε of the sample equals $(5, -5, 5, -5, \ldots)^t$.

In Table 2, we report the average element-wise MSE for the correlation matrix, multiplied by the sample size. The standard errors of the reported results are between 1% and 2% of the respective results. In the case of symmetric contamination, the loss in precision for most estimators due to outliers is relatively small. The MSE of the classical Pearson correlation under 10% contamination is about 60% higher than without outliers. The increase in MSE under contamination is obviously smaller for the other estimators, with the GRCor being the most precise in case of mild contamination (i.e. 1%) and Kendall correlation when higher percentages of the sample are contaminated. In case of asymmetric contamination the results are somewhat different. The precision of the Pearson correlation is now severely affected in the presence of outliers. In case of mild contamination (1%), Kendall has the smallest MSE, closely followed by Spearman. However, for larger percentages of contamination, the MCD correlation estimator has the smallest MSE. The GRCor correlation matrix estimator has larger MSEs under asymmet-

Table 2: Simulated MSE (multiplied by the sample size) of correlation matrix estimators, based on 1000 samples of sizes $n = 50, 100, 200$ from a 10 dimensional multivariate normal distribution with corresponding covariance matrix Σ. All diagonal elements of Σ are equal to 1, all off-diagonal elements equal to 0.5. In the case of symmetric contamination, a fraction ε of the sample follows a multivariate normal distribution with covariance matrix $5 * \Sigma$. For asymmetric contamination, a fraction ε of the sample equals $(5, -5, 5, -5, \ldots)^t$.

			Symmetric contamination			Asymmetric contamination		
n*MSE		$\varepsilon = 0\%$	$\varepsilon = 1\%$	$\varepsilon = 5\%$	$\varepsilon = 10\%$	$\varepsilon = 1\%$	$\varepsilon = 5\%$	$\varepsilon = 10\%$
	Pearson	0.58	0.72	0.87	0.92	8.37	25.35	35.64
	Spearman	0.66	0.71	0.73	0.75	0.89	2.62	5.80
$n = 50$	Kendall	0.67	0.71	0.73	0.75	0.85	2.20	4.83
	GRCor	0.60	0.66	0.73	0.76	1.26	4.23	7.86
	MCD	1.27	1.32	1.28	1.29	1.31	1.30	1.31
	Pearson	0.57	0.65	0.85	0.95	6.08	42.88	70.51
	Spearman	0.67	0.67	0.71	0.75	0.76	3.47	11.00
$n = 100$	Kendall	0.66	0.66	0.70	0.74	0.73	2.81	8.90
	GRCor	0.59	0.61	0.70	0.77	1.05	6.46	15.24
	MCD	1.06	1.02	1.04	1.07	1.01	1.03	0.97
	Pearson	0.56	0.65	0.85	0.99	11.64	84.86	139.52
	Spearman	0.65	0.66	0.71	0.76	0.90	6.28	21.25
$n = 200$	Kendall	0.64	0.65	0.69	0.73	0.82	4.89	16.90
	GRCor	0.57	0.61	0.71	0.81	1.68	12.58	29.80
	MCD	0.81	0.80	0.84	0.84	0.82	0.79	0.81

ric contamination than the other robust estimators we considered, but remains much more robust than the classical estimator.

Positive semidefiniteness: The consistent versions of the Spearman and Kendall correlation matrix estimates are not ensured to be positive semidefinite. We investigate here the severity of this lack of positive definiteness for various levels of outlier contamination and sample sizes. We generate 1000 samples form a multivariate normal distribution where all bivariate correlation coefficients are constant and equal to ρ. As a reference case, we take $\rho = 0.5$, the sample size $n = 60$ and the dimension $d = 40$. Figure 5 plots the percentage of positive semidefinite matrices as estimated by the Spearman (dotted gray) and Kendall (black) correlation estimators for varying values of n and ρ. An important first note is that in case the difference between n and d becomes large (e.g. in case $d = 40$ and $n > 100$) we find that all matrices become positive semidefinite, irrespective of the value of ρ. However, as illustrated on the left graph in Figure 5, for $d = 40$, as the ratio n/d becomes close to 1, the number of positive semidefinite correlation matrix estimates decreases dramatically. Interestingly, this problem is more more severe for Kendall correlation than for Spearman. On the right graph in Figure 5, the percentage of positive semidefinite estimates is plotted against the true bivariate correlation coefficient ρ. This percentage turns out to be quite low for ρ close to zero and increases as ρ approaches 1 for both estimators. Again, we confirm that Kendall correlation yields less positive semidefinite estimates.

Figure 6 plots the percentage of positive semidefinite estimates as a function of the percentage of outliers, for both asymmetric and symmetric contamination. The fraction of outliers barely affects the percentage of positive semidefinite estimates of the Spearman correlation. However, the Kendall correlation suffers substantially

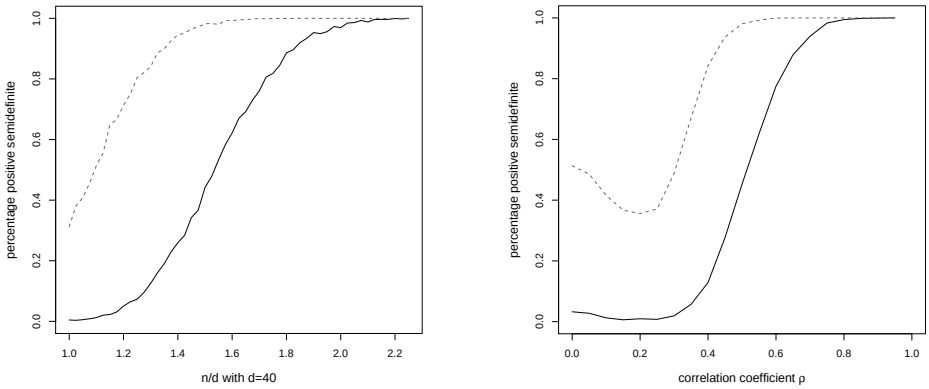

Fig. 5: Percentage of positive semidefinite correlation matrix estimates by Spearman (dotted gray) and Kendall (black) as a function of the ratio n/d (left figure) and as a function of the the correlation coefficient ρ (right figure). The percentages are based on 1000 random samples from a multivariate normal distribution. For the left panel we keep $d = 40$ and $\rho = 0.5$ constant, for the right panel $d = 40$ and $n = 60$.

of a loss in positive semidefinite correlation matrix estimates, as the fraction of outliers increases. In case of asymmetric contamination, 5% of contamination suffices to have that none of the estimated correlation matrices based on the Kendall correlation are positive semidefinite.

5 Application: Robust principal component analysis

The goal of Robust Principal Component Analysis (PCA) is twofold. First, it tries to explain the correlation structure of the data by means of a small number of linear combinations of the original variables, even if there are outliers. Secondly, it allows to flag outliers and to determine of which type they are. We focus on the second application. As shown by Croux and Haesbroeck (2000), among others, robust principal component analysis is easily performed by computing the eigenvalues and eigenvectors of a robust estimator of the correlation matrix. Here, we consider the use of the GRCor in PCA as an alternative to the Pearson correlation. Since we focus solely on the correlation matrix, our analysis starts with robustly standardizing the data using the median and the median absolute deviation (MAD)[1] as measures of location and scale. In what follows, x_i thus denotes the standardized observation, for $1 \leq i \leq n$.

[1] The MAD of a sequence of observations $y_1, \ldots, y_n$ is defined as $1.486\,\text{median}_i|y_i - \text{median}_k y_k|$, where 1.486 is a correction factor such that the MAD is a consistent scale estimator at the normal distribution.

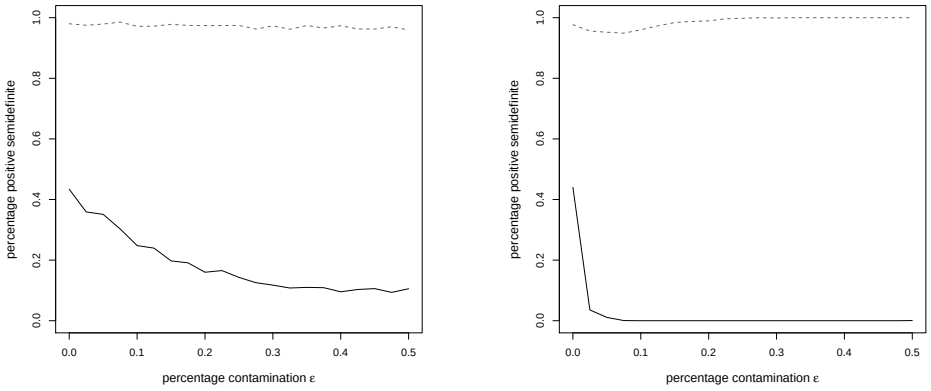

Fig. 6: Percentage of positive semidefinite correlation matrix estimates by Spearman (dotted gray) and Kendall (black) as a function of the percentage of outliers, based on 1000 random samples from a contaminated multivariate normal distribution. We consider random samples with dimension $d = 40$, the sample size $n = 60$ and the bivariate correlation coefficient $\rho = 0.5$. In the left and right panel we consider respectively symmetric and asymmetric outliers.

We consider a dataset containing 8 measurements on 86 containers of milk (Daudin et al., 1988).[2] The eight measurements are: (1) density, (2) fat content, (3) protein content, (4) casein content, (5) cheese dry substance measured in the factory, (6) cheese dry substance measured in the laboratory, (7) milk dry substance and (8) cheese produced, with variables 2-8 measured in grams/liter. Although there are eight measurements, they clearly can be expected to come in groups, making the true dimension probably less. Moreover, previous studies indicated (see e.g. Atkinson et al. (2004)) that this dataset contains several outliers. We investigate whether principal component analysis of the correlation estimator reveals outliers.

To achieve this goal, we use the PCA outlier diagnostic plots of Hubert et al. (2005). They plot the orthogonal distance of each observations versus its score distance. Denote by $t_i = P_k^t x_i$ the score of the i-th observation, with k the number of principal components that is retained and P_k the matrix containing the k eigenvectors corresponding to the k largest eigenvalues in its columns. The score distance of an observation x_i measures the outlyingness of the score vector t_i and is defined as

$$SD_i = \sqrt{\sum_{j=1}^{k} \frac{t_{ij}^2}{l_j}},$$

where l_j are the sorted eigenvalues. The orthogonal distance of x_i measures how far x_i lies from the PCA subspace and is defined as

$$OD_i = ||x_i - P_k t_i||.$$

[2] The dataset is available in the R-package robustbase (Rousseeuw et al., 2009) under the name "milk".

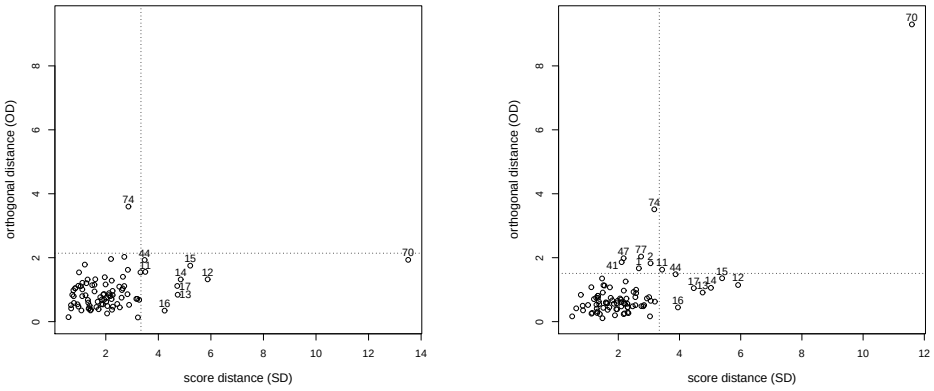

Fig. 7: Diagnostic plot of the contaminated milk dataset based on the Pearson correlation (left) and the GRCor (right).

In Figure 7 the diagnostic plots are shown for the PCA based on the Pearson correlation estimate on the left graph and the GRCor on the right graph. They are based on $k = 4$ retained principal components, since they explain more than 90% of the total variability in both cases. The vertical and horizontal dashed lines indicate the cut-off points for the score and orthogonal distance, which are calculated as in Hubert et al. (2005). Observations that cross one of these lines can be considered as statistical outliers. In the PCA context, three types of outliers can typically be distinguished: (1) bad leverage points correspond to the observations in the upper right quadrant, (2) orthogonal outliers are located in the upper left quadrant and (3) good leverage points in the lower right quadrant.

For PCA based on the Pearson correlation, 10 outliers are detected either as good leverage points or orthogonal outliers. But it is known that outlier detection based on non-robust procedures suffers from the masking effect. Using outlier detection methods based on the forward search algorithm and high breakdown point covariance estimates, Atkinson et al. (2004) found that observation 70 is an extremely pronounced multivariate outlier. The relatively small orthogonal distance of this observation with respect to the Pearson PCA subspace, indicates that it attracted the estimated principal component. In contrast, the orthogonal distance of observation 70 with respect to the GRCor PCA subspace is large, and is correctly flagged as a bad leverage point. The GRCor detects in total 15 outliers, which is in line with the results in Atkinson et al. (2004).

6 Conclusion

In this paper we study the Gaussian rank correlation estimator, which is the usual correlation estimator computed from the normal scores of the data. This type of estimator is known in the literature on rank-based tests, see Hájek and Sidak (1967) for an early reference, and it was already included in the simulation study of Devlin et al. (1975). We propose the GRCor as a valuable robust estimator

for the correlation matrix. It combines several nice properties: (i) it is extremely fast to compute, also in high dimensions (ii) it is always positive semidefinite (iii) it is consistent at multivariate normal distributions (iv) it is quite robust to outliers (v) in absence of outliers, it has 100% asymptotic efficiency at the normal distribution (vi) its definition is very simple. We don't know of any other robust correlation matrix estimator combining these properties. For instance, the OGK-estimator of Maronna and Zamar (2002) lacks properties (v) and (vi). The correlation matrix based on the transformed Spearman or Kendall correlation does not has properties (ii) and (v), and the simulation study in Section 4 showed that non positive semidefinite matrices are frequently obtained, in particular if the number of observations is small with respect to the dimension.

Obviously, the GRCor estimator is not affine equivariant, in contrast to high breakdown covariance matrix estimators as the MCD. In high dimensions, this lack of affine equivariance may become an advantage. If a large number of observations in the data matrix are having aberrant components for only a few variables, then an affine equivariant estimator breaks down, i.e. loses its robustness (see Alqallaf et al. (2009)). The GRCor is much more robust with respect to such "elementwise" contamination in a data matrix.

Although the influence function of the Pearson correlation and the GRCor are the same, their robustness at finite samples is very different. First of all, their sensitivity curves are very different (see Figure 4); the maximum of the sensitivity curve for the Pearson correlation converges at the rate n to infinity, but for the GRCor only at the rate $\log n$. Moreover, the breakdown point as defined in Section 2 is zero for the Pearson correlation, but above 12% for the GRCor. The gain in robustness with respect to the usual correlation matrix estimator has been clearly showed, both theoretically and in the simulation study. There is, however, still room for further improvement on the robustness side, in particular if the percentage of outliers is large.

References

Alqallaf, F., S. Van Aelst, V. Yohai, and R. Zamar (2009). Propagation of outliers in multivariate data. *Annals of Statistics 37*, 311–331.

Alqallaf, F. A., K. P. Konis, R. D. Martin, and R. H. Zamar (2002). Scalable robust covariance and correlation estimates for data mining. In *proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton*.

Atkinson, A. C., M. Riani, and A. Cerioli (2004). *Exploring multivariate data with the forward search*. Springer, Berlin.

Branco, J. A., C. Croux, P. Filzmoser, and M. R. Oliveira (2005). Robust canonical correlations: A comparative study. *Computational Statistics 20*, 203–229.

Capéraà, P. and A. I. G. Guillem (1997). Taux de resistance des tests de rang d'indépendance. *The Canadian Journal of Statistics 25*, 113–124.

Croux, C. and C. Dehon (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods and Applications*, forthcoming.

Croux, C. and G. Haesbroeck (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika 87*, 603–618.

Daudin, J. J., C. Duby, and P. Trecourt (1988). Stability of principal component analysis studied by the bootstrap method. *Statistics 19*, 241–258.

Davies, P. and U. Gather (2005). Breakdown and groups (with discussion). *Annals of Statistics 33*, 977–1035.

Devlin, S., R. Gnanadesikan, and J. Kettering (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika 62*, 531–545.

Dominici, D. E. (2003). The inverse of the cumulative standard normal probability function. *Integral Transforms and Special Functions 14*, 281–292.

Grize, Y. (1978). *Robustheitseigenschaften von Korrelations-schätzungen*. Ph. D. thesis, ETH Zürich.

Hájek, J. and Z. Sidak (1967). *Theory of Rank Tests*. Academic Press, New York.

Hubert, M., P. Rousseeuw, and K. Vanden Branden (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics 47*, 64–79.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika 30*, 81–93.

Khan, J., S. Van Aelst, and R. Zamar (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association 480*, 1289–1299.

Maronna, R. and R. Zamar (2002). Robust estimates of location and dispersion of high-dimensional datasets. *Technometrics 44*, 307–317.

Maronna, R. A., R. D. Martin, and V. J. Yohai (2006). *Robust statistics: theory and methods*. John Wiley & Sons Ltd, England.

Rousseeuw, P., C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, and M. Maechler (2009). *robustbase: Basic Robust Statistics*. R package version 0.5-0-1. http://CRAN.R-project.org/package=robustbase.

Rousseeuw, P. and K. Van Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41*, 212–223.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology 15*, 201–293.

Van Aelst, S., E. Vandervieren, and G. Willems (2010). Robust principal component analysis based on pairwise correlation estimators. In *Proceedings of the 19th International Conference on Computational Statistics, Paris*.

A Proofs

Proof of Proposition 2. Here, we focus on the situation where k is even, an analogous proof can be given for odd k. Using symmetry, we can rewrite (2.2) as

$$\hat{\rho}(S_n^{k^*}) = -4\frac{\frac{1}{n}\sum_{i=1}^{k/2} z_i^2}{\frac{1}{n}\sum_{i=1}^{n} z_i^2} + 1, \tag{A.1}$$

with n the sample size and k the (even) number of contaminated elements.

Write $k = \varepsilon n$, with $0 < \varepsilon < 1$. We need to find the smallest ε such that the correlation estimate (A.1) becomes negative. For n tending to infinity, the condition that (A.1) is negative becomes

$$-4\int_{-\infty}^{\Phi^{-1}(\varepsilon/2)} z^2\phi(z)dz + 1 \leq 0. \tag{A.2}$$

The asymptotic breakdown point ε^* is the smallest ε such that (A.2) holds. Note that the left hand side of (A.2) is strictly decreasing in ε, equals -1 for $\varepsilon = 1$ and 1 for $\varepsilon = 0$. So ε^* is the

unique solution of

$$\int_{-\infty}^{\Phi^{-1}(\varepsilon/2)} z^2\phi(z)dz = \frac{1}{4}. \tag{A.3}$$

Finally, note that (A.3) can be rewritten as (2.3) using partial integration.

Proof of Proposition 3. Recall that $S_{n-1} = \{(x_1, x_1), \ldots, (x_{n-1}, x_{n-1})\}$. It follows from Proposition 1, with $k = 1$, that the bias induced by adding one observation is maximal when adding the couple (x, y) for which $x < \min_i x_i$ and $y < \max_i y_i$ for $i = 1, \ldots, n-1$. From Equation (2.2), with $k = 1$, it then follows that the maximum of the sensitivity curve, as defined in (3.1), is given by

$$\gamma_n^* = |n * (c_n \sum_{i=2}^{n} (z_i z_{i-1} - z_1^2) - 1)|,$$

with $c_n = 1/\sum_{i=1}^n z_i^2$ and $z_i = \Phi^{-1}(\frac{i}{n+1})$. For $n \to \infty$, $(nc_n) \xrightarrow{p} \frac{1}{E[Z^2]} = 1$ and $\frac{1}{n}\sum_{i=2}^n z_i z_{i-1} \xrightarrow{p} E[Z^2] = 1$, with $Z \sim N(0,1)$. We therefore have that,

$$\gamma_n^* \sim n - n(1 - \frac{z_1^2}{n}) = z_1^2 = \Phi^{-1}(\frac{1}{n+1})^2.$$

Proposition 21 of Dominici (2003) states that

$$\Phi^{-1}(x) \sim -\sqrt{LW(\frac{1}{2\pi x^2})},\ x \to 0, \tag{A.4}$$

where $LW(x)$ is the Lambert W function, defined by the implicit equation $LW(x)\exp(LW(x)) = x$, having asymptotic behavior $LW(x) \sim \ln(x) - \ln(\ln(x))$, for $x \to \infty$. We conclude that

$$\gamma^* \sim LW(\frac{(n+1)^2}{2\pi}) \sim 2\log(n),$$

for $n \to \infty$.

Proof of Proposition 4. Define $H_\varepsilon = (1-\varepsilon)H + \varepsilon\Delta_{(x_0,y_0)}$, where H is a bivariate normal distribution with correlation ρ. It follows from (3.3) that

$$\mathrm{GRCor}(H_\varepsilon) = \varepsilon h(F_\varepsilon(x_0))h(G_\varepsilon(y_0)) + (1-\varepsilon)\int h(F_\varepsilon(x))h(G_\varepsilon(y))dH(x,y), \tag{A.5}$$

with $F_\varepsilon(x) = (1-\varepsilon)F(x) + \varepsilon I(x \geq x_0)$, $G_\varepsilon(x) = (1-\varepsilon)G(x) + \varepsilon I(x \geq x_0)$, and $h = \Phi^{-1}$. At the model distribution $H = \Phi_\rho$, we have that $F = G = \Phi$. Computing the derivative of (A.5) and evaluating at $\varepsilon = 0$ yields the influence function

$$\begin{aligned}
&\mathrm{IF}((x_0,y_0), \mathrm{GRCor}, H)\\
&= h(\Phi(x_0))h(\Phi(y_0)) - \rho\\
&+ \int h'(F_\varepsilon(x))\frac{\delta}{\delta\varepsilon}F_\varepsilon(x)|_{\varepsilon=0}\, y\, dH(x,y) + \int h'(G_\varepsilon(y))\frac{\delta}{\delta\varepsilon}G_\varepsilon(y)|_{\varepsilon=0}\, x\, dH(x,y)\\
&= x_0y_0 - \rho\\
&+ E[h'(\Phi(X))\{-\Phi(X) + I(X \geq x_o)\}Y] + E[h'(\Phi(Y))\{-\Phi(X) + I(Y \geq y_o)\}X],
\end{aligned} \tag{A.6}$$

where the expectation is with respect to H. Note the similarity in the last two terms of (A.6). Due to bivariate normality, we may write write $Y = \rho X + \sqrt{1-\rho^2}\,\varepsilon$ with ε independent of

X. The third term of (A.6) then becomes

$$\begin{aligned}
&E[h'(\Phi(X))\{-\Phi(X)+I(X\geq x_o)\}Y]\\
&=\rho E[\frac{X}{\phi(X)}\{I(X\geq x_0)-\Phi(X)\}]\\
&=\rho\int\frac{x}{\phi(x)}\{I(x\geq x_0)-\Phi(x)\}\phi(x)dx\\
&=\rho\lim_{M_1,M_2\to\infty}\int_{-M_1}^{M_2}x\{I(x\geq x_0)-\Phi(x)\}dx\\
&=\rho\lim_{M_1,M_2\to\infty}\int_{x_0}^{M2}xdx-\int_{-M_1}^{M_2}x\Phi(x)dx\\
&=\rho\lim_{M_1,M_2\to\infty}\{\frac{M_2^2}{2}-\frac{x_0^2}{2}-\frac{x^2}{2}\Phi(x)\mid_{-M_1}^{M_2}+\int_{-M_1}^{M_2}\frac{x^2}{2}d\Phi(x)\}\\
&=\rho[\frac{-x_0^2}{2}+\int_{-\infty}^{\infty}\frac{x^2}{2}d\Phi(x)+\lim_{M_1,M_2\to\infty}\{\frac{-M_1^2}{2}\Phi(-M_1)+\frac{M_2^2}{2}(1-\Phi(M_2))\}]\\
&=\rho[\frac{-x_0^2}{2}+\frac{1}{2}].
\end{aligned}\tag{A.7}$$

The last term of (A.6) can be simplified in a similar way and (A.6) then simplifies to

$$\mathrm{IF}((x_0,y_0),\mathrm{GRCor},\Phi_\rho)=x_0y_0-\rho+\rho(\frac{-x_0^2+1}{2})+\rho(\frac{-y_0^2+1}{2}),\tag{A.8}$$

resulting in (3.4).